Purifying selection acts on coding and non-coding sequences of paralogous genes in Arabidopsis thaliana

Hoffmann, Robert D; Palmgren, Michael Broberg

Published in:
B M C Genomics

DOI:
10.1186/s12864-016-2803-2

Publication date:
2016

Document version
Publisher's PDF, also known as Version of record

Document license:
Unspecified

Citation for published version (APA):
Hoffmann, R. D., & Palmgren, M. B. (2016). Purifying selection acts on coding and non-coding sequences of paralogous genes in *Arabidopsis thaliana*. *B M C Genomics*, *17*, [456]. https://doi.org/10.1186/s12864-016-2803-2

Hoffmann and Palmgren *BMC Genomics* (2016) 17:456
DOI 10.1186/s12864-016-2803-2

BMC Genomics

RESEARCH ARTICLE Open Access

Purifying selection acts on coding and non-coding sequences of paralogous genes in *Arabidopsis thaliana*

Robert D. Hoffmann* and Michael Palmgren

Abstract

Background: Whole-genome duplications in the ancestors of many diverse species provided the genetic material for evolutionary novelty. Several models explain the retention of paralogous genes. However, how these models are reflected in the evolution of coding and non-coding sequences of paralogous genes is unknown.

Results: Here, we analyzed the coding and non-coding sequences of paralogous genes in *Arabidopsis thaliana* and compared these sequences with those of orthologous genes in *Arabidopsis lyrata*. Paralogs with lower expression than their duplicate had more nonsynonymous substitutions, were more likely to fractionate, and exhibited less similar expression patterns with their orthologs in the other species. Also, lower-expressed genes had greater tissue specificity. Orthologous conserved non-coding sequences in the promoters, introns, and 3′ untranslated regions were less abundant at lower-expressed genes compared to their higher-expressed paralogs. A gene ontology (GO) term enrichment analysis showed that paralogs with similar expression levels were enriched in GO terms related to ribosomes, whereas paralogs with different expression levels were enriched in terms associated with stress responses.

Conclusions: Loss of conserved non-coding sequences in one gene of a paralogous gene pair correlates with reduced expression levels that are more tissue specific. Together with increased mutation rates in the coding sequences, this suggests that similar forces of purifying selection act on coding and non-coding sequences. We propose that coding and non-coding sequences evolve concurrently following gene duplication.

Keywords: *Arabidopsis thaliana*, Conserved non-coding sequences, Evolution, Gene expression, mRNA, Paralogous genes, Promoters, 3′ UTR

Background

A striking difference between metazoans and plants is the recent occurrence of whole genome duplication (WGD) events in plants [1–3]. At least three WGD events have been confirmed in the ancestry of *Arabidopsis thaliana* [4, 5], with the most recent one (entitled alpha; α) occurring around 23 Mya [4]. After a polyploidization event, the genome reorganizes and, although many duplicated sequences are deleted, a considerable proportion of duplicated genes remains as paralogs in the genome [1]. *A. thaliana* contains more than 2500 paralogous gene pairs, accounting for about one-sixth of all protein-coding genes in this species [1, 6]. Due to the wealth of paralogous gene pairs arising from WGD and the reduced selection pressure on redundant gene copies, WGD is thought to provide the potential for adaptive radiation and evolutionary innovations [7–10].

Several models of evolution following a WGD event have been proposed, the most prominent of which are balanced gene drive [11], subfunctionalization of gene pairs [12], and neofunctionalization [9, 13] (reviewed in [14]). The balanced gene drive model is based on the gene balance hypothesis, which predicts that duplicates are retained when the duplication leads to a new balance between the products of dosage-dependent genes [15]. For instance, when the proteins encoded by paralogous genes function as part of a protein complex, the loss of one paralog would change the strength or nature of interactions in the complex, and therefore both copies are likely to be retained [11, 16]. Subfunctionalization

* Correspondence: hoffmann@plen.ku.dk
Center for Membrane Pumps in Cells and Disease - PUMPKIN, Danish National Research Foundation, Department of Plant and Environmental Sciences, University of Copenhagen, 1871 Frederiksberg C, Denmark

describes the process of dividing an ancestral gene function between the two members of a paralogous gene pair. Accordingly, fulfilling the ancestral function now requires duplicate genes [17]. Mutations that lead to new functions of duplicated genes can occur in both protein-coding and non-coding regions [9, 18, 19], and the functional classes of paralogs are suggested to be linked to gene expression [20].

As predicted by the balanced gene drive model, genes encoding subunits of protein complexes or enzymes of the same metabolic pathway tend to be retained after WGD, as shown in ciliates [21, 22], yeast [23], and plants [24]. Genes involved in developmental processes, regulation of transcription, and signal transduction are preferentially retained as duplicates [18, 25–28]. These functional categories suggest that neo-/subfunctionalization drive retention of the duplicates. Stress-responsive genes were found to be retained after WGD, suggesting that environmental challenges promote biased duplicate retention [29]. When paralogs were separated into pairs with similar or differential expression, it was found that DNA- and nucleic acid-binding were overrepresented among similarly expressed paralog pairs, while functions related to biosynthesis and metabolism were overrepresented among the differentially expressed pairs [20]. During the course of evolution, paralogs diverge in amino acid sequence [21] and gene expression profile [18]. Furthermore, paralog coexpression correlates with the number of shared regulatory motifs [30–32]. However, how the coding and non-coding sequences of the same paralog evolve is unknown.

Orthologous conserved non-coding sequences (CNSs) are a characteristic of eukaryotic genomes. As these sequences of non-coding DNA are evolutionarily conserved across species [33, 34], they are thought to have a biological function [35]. Such CNSs are located in introns, intergenic regions proximal or distal to genes, and the 5′ and 3′ untranslated regions (UTRs) of genes. Comparative genomic studies have identified thousands of CNSs in the genomes of humans and model organisms such as mouse and *A. thaliana* [36–41]. In plants, CNSs have been hypothesized to affect the transcription levels of neighboring genes [33, 42] and several studies have shown that CNSs are enriched for transcription factor binding sites [36, 40, 43–45]. Published CNS datasets often overlap to a limited degree only [46], depending on the included species and the detection parameters. Four studies report the identification of CNSs using *A. thaliana* as a reference [36, 38–40]. Baxter et al. (2012) aligned four dicot species and reported 1865 CNSs in the region upstream of the transcription start site (TSS) of 1643 genes [36]. Hupalo and Kern (2013) identified CNSs in 20 angiosperm species using deep whole-genome alignment [39]. Haudry et al. (2013) aligned the genomes of nine members of the Brassicaceae family, identifying over 90,000 CNSs [38]. Van de Velde et al. (2014) used phylogenetic footprinting and whole-genome alignment to identify CNSs in 12 dicot species [40].

In this study, we aimed to identify differences between WGD-derived paralog pairs with similar expression levels and those with different expression levels. We found that the genes of differentially expressed paralog pairs with reduced expression are under less purifying selection, exhibit more tissue-specific expression, and have lost orthologous CNSs compared with paralogs that have equal or increased expression. Paralog pairs with similar expression strength may be retained by gene-dosage constraints, while neo- and/or subfunctionalization may drive retention of differentially expressed pairs.

Results

Classification of paralogous genes based on expression levels

To estimate the average gene transcript levels in *A. thaliana*, we averaged the expression levels of 19,765 genes in 15 tissues or cell types [47] (Additional file 1: Table S1). Among these genes, we identified a set of 1312 highly similar paralogs resulting from the recent α-WGD event in *A. thaliana* (Additional file 1: Table S2). We classified paralogous pairs based on their relative expression levels, and grouped together 245 and 156 pairs with similar ($\log_2$-ratio < 1.25) and differential ($\log_2$-ratio ≥ 7) expression levels, respectively. Paralog pairs with differential expression were further categorized into those with higher expression and lower expression (Fig. 1). To compare paralogs that exhibited similar expression with those that had higher and lower expression, one gene of every similarly expressed pair was randomly selected (Additional file 1: Tables S2 and S3), and these randomly selected genes were used in further analyses.

Differentially expressed paralogs are subject to relaxed purifying selection

To characterize the *A. thaliana* genes in each group (i.e., similar expression, higher expression, and lower expression groups), we compared their expression with that of their orthologs in *Arabidopsis lyrata*. The two species diverged after the α-WGD event and the *A. lyrata* genome has been sequenced [48] and transcript level data are available [49].

Firstly, we analyzed how many genes of the higher- and lower-expressed paralogs have orthologs in *A. lyrata*. Amongst the 156 paralogous pairs with differential expression levels, the lower-expressed parologs had 144 orthologs, whereas the higher-expressed paralogs had 155 orthologs ($P = 0.0028$, Fisher's exact test, two-sided) (Fig. 2a).

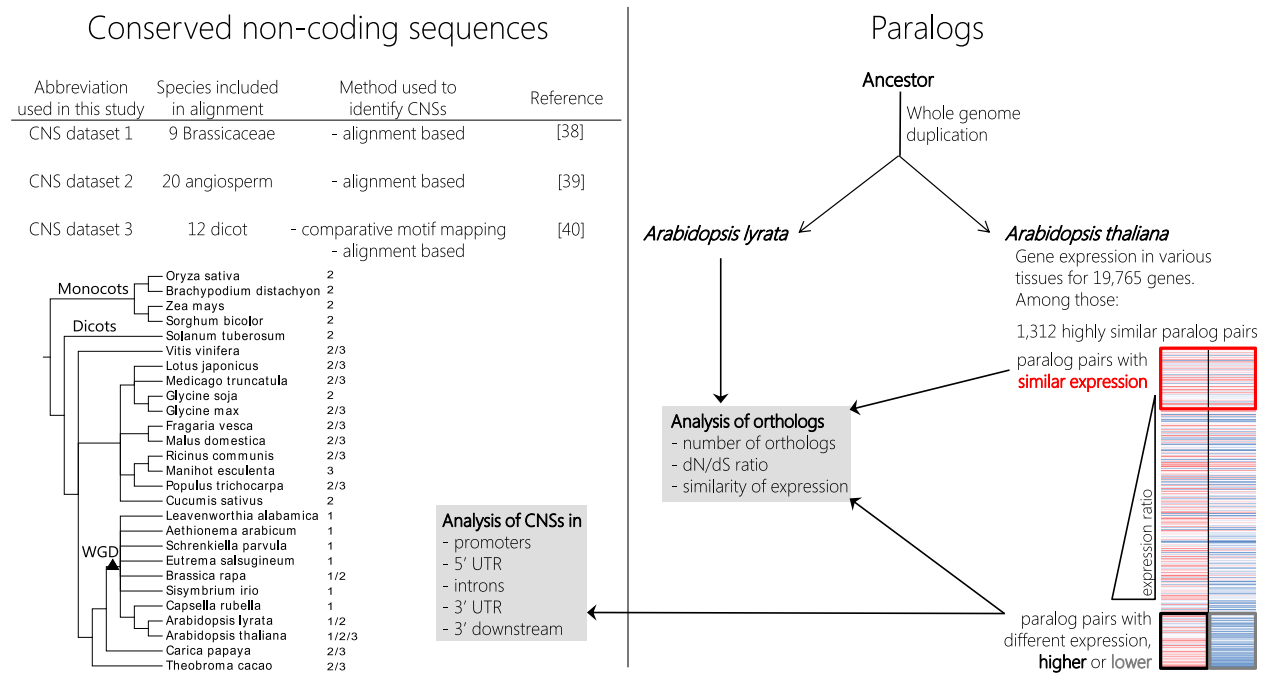


Fig. 1 Flowchart of the analyses described in this paper. Based on their relative expression levels (red, higher expression; blue, lower expression), paralogous pairs were grouped into those having similar (red box; $\log_2$-ratio < 1.25) or differential ($\log_2$-ratio ≥ 7) expression. Differentially expressed paralogs were further grouped into "higher expression" (black box) or "lower expression" (gray box). Conserved non-coding *A. thaliana* sequences reported in three published studies were analyzed for their presence at paralogous gene pairs

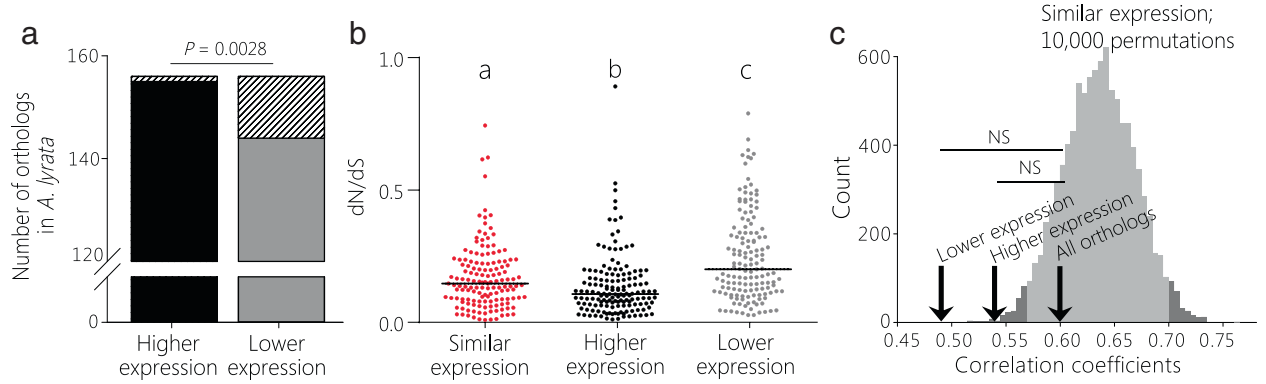


Fig. 2 Paralogs with lower expression experience relaxed purifying selection. **a** Paralogs in the higher expression group (black bars) have more orthologs in *A. lyrata* than do their paralogs with lower expression (grey bars) (Fisher's exact test, two-sided). White hatched bars show number of genes in *A. thaliana* that do not have an ortholog in *A. lyrata*. **b** The dN/dS ratio of *A. thaliana* and *A. lyrata* orthologs was compared between paralogs with similar expression (one paralog of each pair) and paralogs in groups with higher expression and lower expression. Letters indicate significant differences in sample distribution ($P < 0.001$, Kruskal-Wallis test with Dunn's correction). **c** The gene expression profiles of paralogs with lower expression are less correlated with those of *A. lyrata* orthologs than are paralogs with higher or similar expression. Spearman correlation coefficients based on expression in *A. thaliana* and *A. lyrata* were calculated for paralogs with lower expression ($\rho = 0.49$, $n = 109$), higher expression ($\rho = 0.54$, $n = 147$), and all orthologs ($\rho = 0.6$, $n = 15{,}134$), indicated by black arrows. The correlation coefficient for differentially-expressed paralogs is different from that for paralogs with similar expression; grey bars show the distribution of ρ values based on 10,000 permutated sets of paralogs from pairs with similar expression ($n = 245$). Light grey bars indicate the distribution of 95 % of the resulting correlation coefficients, and dark grey bars indicate the 2.5 % distribution on either side. For details of the statistical analysis see main text

Next, to estimate the evolutionary rate of divergence of the paralogous genes' DNA sequences, we analyzed synonymous (dS) and non-synonymous (dN) substitution rates for *A. thaliana* and *A. lyrata* orthologs. A lower dN/dS ratio is indicative of purifying selection. We detected dN/dS ratios of 0.17, 0.13, and 0.24 for genes in the groups with similar, higher, and lower expression, respectively ($P < 0.001$, Kruskal-Wallis test with Dunn's correction for multiple testing) (Fig. 2b).

Lastly, we compared gene expression levels using data sampled in flowers at stages 1-14 (*A. lyrata*) and total inflorescences containing flowers at stages 1-14 *(A. thaliana*) [49] (Additional file 1: Table S4). We calculated Spearman correlation coefficients (ρ) for all orthologs ($n = 15{,}134$, $\rho = 0.6$, $P < 0.0001$) and paralogs with higher ($n = 147$, $\rho = 0.54$, $P < 0.0001$) and lower expression ($n = 109$, $\rho = 0.49$, $P < 0.0001$) (Fig. 2c; Additional file 2: Figure S1). Expression correlation between lower- or higher-expressed paralogs and all orthologs was not significantly different (Fig. 2c). We performed the same analysis with ortholog expression data from four species of the Brassicaceae family (*A. thaliana*, *A. lyrata*, *Capsella rubella*, and *Capsella grandiflora*) [50]. The results also showed reduced expression correlation of the lower-expressed paralogs compared to the higher-expressed paralogs (Additional file 2: Figure S2). To further test differences in correlation of gene expression between *A. thaliana* and *A. lyrata* orthologs, we approximated the median correlation co-efficient for the similarly expressed paralogs. As the expression correlation for the similarly expressed paralogs would be different each time we randomly selected one gene from each pair, we computed the correlation coefficients for 10,000 repeated random selections. The resulting correlations had a median ρ of 0.64 (with 95 % of values between 0.57 and 0.7), thus overlapping ρ of all orthologs, but not those with differential expression (Fig. 2c; Additional file 1: Table S5). Taken together, these findings suggest that the lower-expressed gene of a paralogous pair experiences reduced purifying selection.

Paralogous genes with different expression levels function in stress responses

We examined the functions of genes in groups with similar and differential expression by performing a gene ontology (GO) term enrichment analysis (Table 1; see Additional file 1: Table S6 for a full list of enriched GO terms). Paralogs with similar average expression levels were found to be enriched for being components of ribosomes, whereas those with differential expression levels were enriched for responses to different abiotic and biotic stresses (Table 1).

We reasoned that the stress response is often a rather local action, restricted to certain tissues. Tissue specificity can be measured with the index τ [51], for which values approaching 0 indicate broad gene expression and those approaching 1 indicate tissue-specific expression. We found that expression maxima for 22 % of differentially expressed paralogs were in the same tissue/cell type for both genes (Additional file 1: Table S7). Interestingly,

Table 1 GO term enrichment of paralogous genes with similar or different expression levels

GO term category[c]	Differentially (ratio ≥ 7) expressed[a]			Similarly (ratio < 1.25) expressed[a]		
	FDR[b]	Subset ratio	GO term	FDR[b]	Subset ratio	GO term
Molecular function	0.0230	50 %	catalytic activity			
	0.0360	20 %	hydrolase activity			
	0.0083	14 %	transporter activity			
	0.0100	10 %	substrate-specific transporter activity			
	0.0230	10 %	transmembrane transporter activity			
Biological process	0.0430	48 %	cellular process			
	0.0006	29 %	response to stimulus			
	0.0190	23 %	cellular biosynthetic process			
	0.0330	23 %	biosynthetic process			
	0.0007	19 %	response to stress			
Cellular component	0.0110	36 %	cytoplasm	0.0390	10 %	plasma membrane
	0.0087	35 %	cytoplasmic part	0.0180	7 %	cytosol
	0.0098	26 %	membrane	0.0330	5 %	ribonucleoprotein complex
	0.0017	16 %	plasma membrane	0.0120	4 %	cytosolic part
	0.0110	36 %	cytosol	0.0300	4 %	ribosome

[a]For each pair, the gene with higher expression was selected for analysis
[b]False Discovery Rate (FDR)
[c]For each GO term category, the five entries with the highest Subset ratio are shown

we found that paralogous genes with similar average expression levels had strongly correlating degrees of tissue specificity, whereas the lower-expressed genes of differentially expressed paralogous pairs were skewed towards tissue-specific expression (Fig. 3). This finding suggests that members of differentially expressed paralogous pairs did not simply experience a reduction in average gene expression, but gained tissue-specific expression.

Number of CNSs in promoter regions, introns, and 3′ UTRs correlates with relative gene expression

Since CNSs are indicative of *cis*-elements that regulate transcription, we compared the number of CNSs associated with paralogous genes in each group. For paralogs with higher expression, no significant correlation was observed between the number of CNSs in the promoter regions and the paralogous pair expression ratio (i.e., the ratio of paralog expression levels; Fig. 4; Additional file 1: Table S8). By contrast, for lower-expressed parolgs, we found a significant negative correlation between the number of CNSs in the promoter regions and diverging paralog expression levels. The correlation was strongest for CNSs of dataset 1 (Kendall's *tau-b* = 0.11, $P < 0.0001$), but was observed for CNSs obtained from all three CNS datasets (Fig. 4).

Similar results were obtained for CNSs in introns and 3′ untranslated regions (3′ UTRs). In CNS dataset 1, we detected a negative correlation between the paralogous pair expression ratio and the number of CNSs in introns from the lower-expressed gene (Kendall's *tau-b* = -0.07, $P = 0.002$) but not the number of CNSs from the higher-expressed gene (Kendall's *tau*-b = 0.0, $P = 0.9$) (Fig. 4; Additional file 1: Table S9). CNSs from datasets 2 and 3 (which are from phylogenetically more distant dicots or angiosperms) showed no significant correlation with gene expression. For paralog expression ratio and CNSs in 3′ UTRs, we found negative correlations between the expression of the lower-expressed gene of a paralogous pair and the number of CNSs in CNS dataset 1 (Kendall's *tau-b* = -0.14, $P < 0.0001$) and CNS dataset 2 (Kendall's *tau-b* = -0.06, $P = 0.01$) (Fig. 4). As with CNSs in the promoter regions and introns, higher-expressed paralogs were not significantly correlated with the number of 3′ UTR CNSs.

It is possible that large stretches of conserved sequence in the genome of a common ancestor were fragmented into several shorter CNSs [52, 53], resulting in a miscalculation of the number of CNSs present today. Therefore, we tested the correlation between divergence in gene expression and the sum of bases that form CNSs at any given gene. The results are similar to those obtained in the analysis of single-element CNSs (Additional file 2: Figure S3), suggesting that CNSs likely are not fragments of elements that were previously larger.

In conclusion, we observed that an increase in differential gene expression level is negatively correlated with the number of CNSs located 5′ upstream, in introns, and in 3′ UTRs of the lower-expressed member of the pair.

CNSs at similarly and differentially expressed genes have comparable properties

We next analyzed if transcription factor (TF) binding motifs were enriched or depleted in CNSs located in promoter regions. For this, we mapped the positions of binding motifs from 274 TFs, belonging to 30 families,

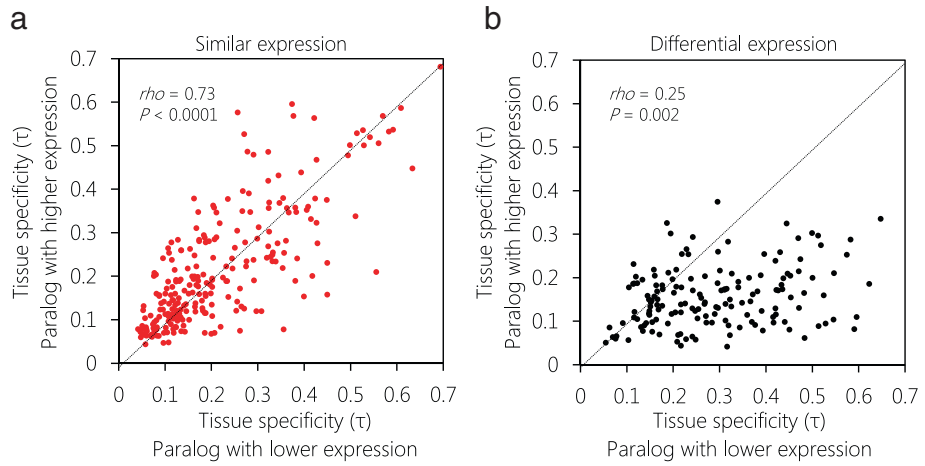


Fig. 3 Differentially expressed paralogous pairs diverge in tissue specificity. Paralogous pairs were selected for similar (log_2-ratio < 1.25, $n = 245$) or differential (log_2-ratio ≥ 7, $n = 156$) average expression levels. **a** Tissue specificity (τ) is correlated for the two genes of a pair with similar expression levels. **b** Lower-expressed genes of pairs with differential gene expression have stronger tissue specificity than do their higher-expressed paralogs

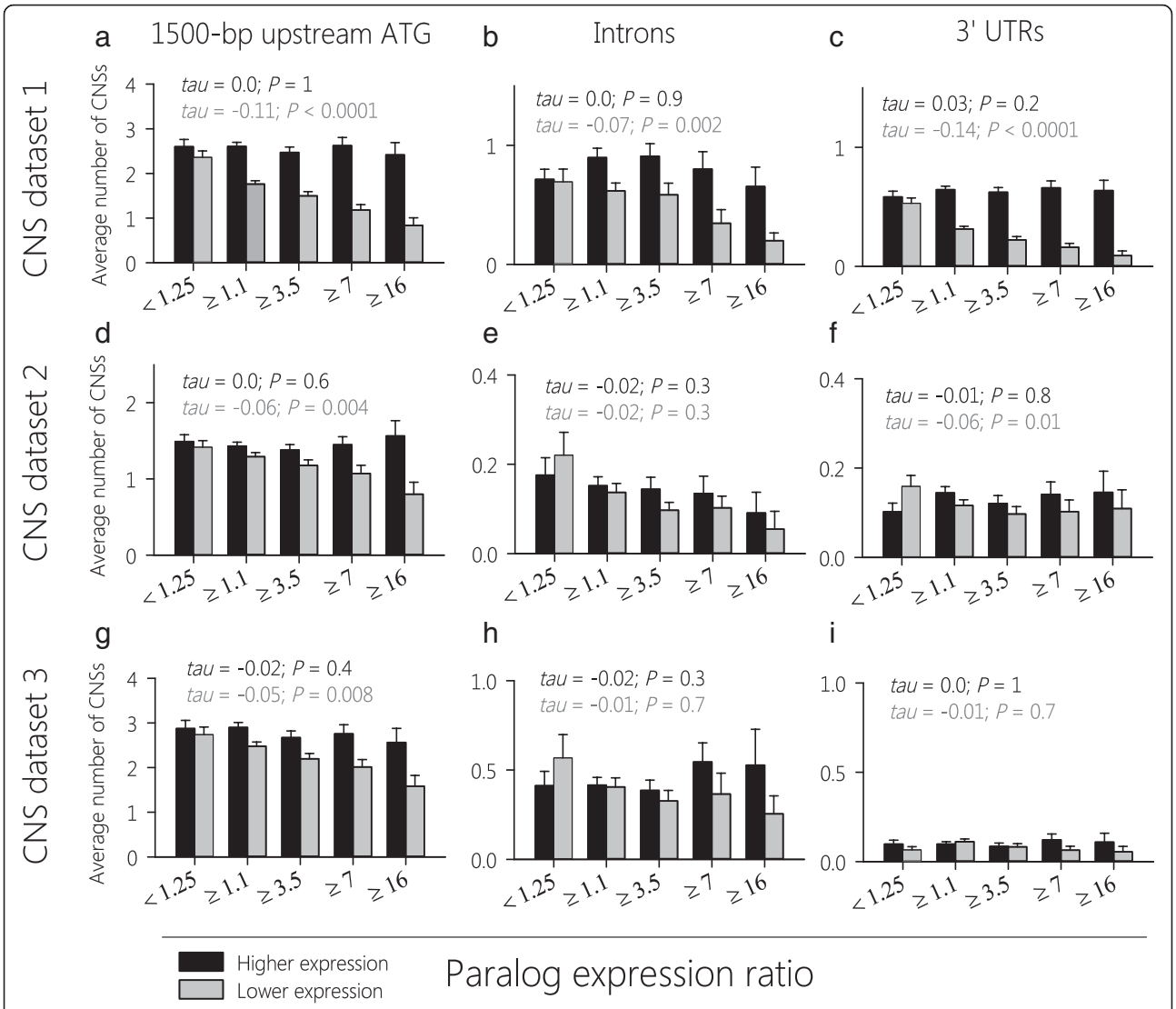
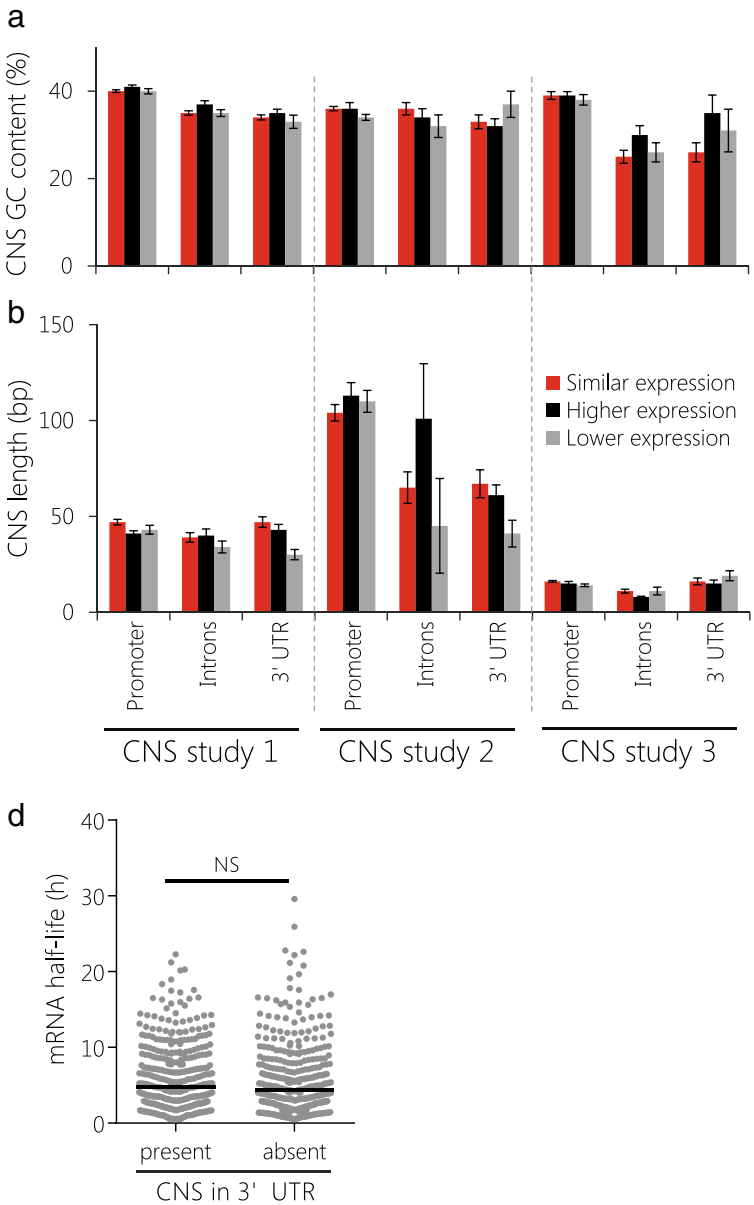


Fig. 4 With increasing expression divergence, lower-expressed paralogs have fewer CNSs in their promoters, introns, and 3′ UTRs. The two genes of each paralogous pair were classified as higher-expressed (black bars) or lower-expressed (grey bars) paralog, and plotted according to their ratio in gene expression levels (based on $\log_2$-values). Y-axes show the average number of CNSs within promoters (1500-bp upstream ATG -100 bp) (**a**, **d**, **g**), introns (**b**, **e**, **h**), and 3′ UTRs (**c**, **f**, **i**). Three different CNS datasets were analyzed [38–40]. A significant negative correlation between the paralog expression ratio and number of CNSs in promoters (Kendall rank correlation) was found for the lower expression group in all datasets. A correlation between paralog expression ratio and number of CNSs in introns or 3′ UTRs was limited (**b**, **c**, **f**). The higher expression group had no significant correlation with the number of CNSs. Error bars are given as s.e.m

within CNSs from CNS dataset 1 (since this dataset is based on the most closely related species). We found that TF binding motifs of several families were enriched or depleted in CNSs present at paralogs in the groups of similar, higher, or differential expression. However, from 90 measurements, only two showed statistically significant deviations ($P < 0.01$; resampling without replacement; 1000 iterations) (Fig. 5c). Paralogs with higher expression are depleted (0.6 fold) of MYB TF binding motifs, and similarly expressed paralogs are depleted (0.5 fold) of HD-ZIP TF binding motifs.

Having identified a negative correlation between paralog expression ratio and 3′ UTR CNSs, we examined whether 3′ UTR CNSs (identified from CNS dataset 1) influence mRNA stability. For this, we compared the mRNA half-life data of paralogous pairs that had one paralog that lacked CNSs in the 3′ UTR with those of paralogous pairs that had one paralog with one or more 3′ UTR CNSs ($n = 365$, Fig. 5d). No significant difference in the corresponding mRNA half-lives was found (Wilcoxon signed-rank test, $P = 0.15$) (Additional file 1: Tables S10 and S11).



c

Transcription factor family	Similar expression	Higher expression	Lower expression
ABI3/VP1	0.6	0.6	1.7
AP2/EREBP	0.9	0.8	0.7
AT-Hook	1.0	1.0	0.9
bHLH	1.6	1.4	0.2
bZIP	1.1	0.8	0.1
C2C2(Zn) Dof	1.1	0.8	0.6
C2C2(Zn) GATA	1.3	0.6	1.1
C2C2(Zn) YABBY	0.8	0.8	0.6
C2H2(Zn)	0.6	0.5	0.4
CAMTA	1.9		
CAMTA-like		1.6	
E2F/DP			
GARP	1.0	0.5	1.1
GARP/ARR-B	1.5	1.1	1.4
GATA	1.7	0.5	0.3
GRF	0.3		
HD-HOX	1.0	1.0	0.6
HD-KNOTTED	1.1	0.6	0.2
HD-PHD	0.8	0.9	1.8
HD-ZIP	**0.5**	1.1	0.7
HSF	1.0	0.3	1.1
LFY	1.1	0.6	1.0
LOB/AS2	1.7	0.5	1.3
MADS	2.0	0.5	
MYB	0.9	**0.6**	1.2
NAC	1.1	1.1	0.6
SBP	1.3	0.8	0.4
SRS	0.8		
Trihelix	0.9	0.5	0.8
WRKY(Zn)	1.1	0.6	1.9

Fig. 5 Length, GC content, transcription factor binding sites, and effect on mRNA half-life are similar for CNSs at paralogs. GC content (**a**) and length (**b**) of CNSs from three datasets are compared between paralogs with similar (red), higher (black), or lower (grey) expression in promoters, intronic regions, and 3′ UTRs. No statistically significant differences were found (resampling without replacement, 1000 iterations). Error bars are given as s.e.m. **c** Transcription factor binding site enrichment (red) and depletion (blue) within CNSs located in promoter regions is shown. CNSs are from CNS dataset 1. Missing values (grey) indicate complete absence of the transcription factor binding site motif from CNSs in the respective paralog group. Significant depletion was found for MYB binding site motifs in the higher expression group and HD-ZIP binding site motifs in the similar expression group (resampling without replacement, 1000 iterations; $P < 0.01$). **d** Paralogs with no (absent) or at least one (present) CNS in their 3′ UTRs show no difference in mRNA half-lives (Wilcoxon signed-rank test; black line indicates median value). NS = not significant

We lastly compared the average length and GC content of CNSs from all three CNS datasets in the promoter regions, introns, and 3′ UTRs (Fig. 5a and b). For paralogs in the groups of similar, higher, and lower expression, differences in either CNS length or GC content were small and not statistically significant ($P < 0.01$; resampling without replacement; 1000 iterations).

Discussion

The aim of this study was to elucidate if and how the retention of paralogs is connected to gene expression.

The results provide evidence for a concurrent purifying selection on coding and noncoding sequences of paralogous genes in *A. thaliana*.

Mutation rates in coding sequences

In our study, we separated two paralogous genes based on their expression strength and measured dN/dS values by comparing each gene with its ortholog in the *A. lyrata* genome. This allowed us to gauge the evolutionary rate of two paralogous genes individually. We found that the lower-expressed genes have acquired more nonsynonymous mutations. This is supported by the finding that paralogs with more similar expression have lower sequence diversity than do paralogs with differential expression [20]. Also, lower-expressed paralogs are more likely to be lost in *A. lyrata*, suggesting that the lost orthologs were under neutral selection. Considering all this, the results suggest that the lower-expressed paralogs are under less purifying selection than are the higher- and similarly expressed paralogs.

Mutation rates in non-coding sequences

Paralogous genes with similar expression strength were found to have similar tissue specificity. Genes of differentially expressed pairs, however, had diverged in tissue specificity. Higher-expressed genes were broadly expressed, whereas their lower-expressed paralogs were more tissue specific, which is similar to the expression pattern reported for paralogs functioning in stress responses [54]. In the gene expression correlation analysis of *A. thaliana* paralogs and their Brassicaceae orthologs, we found that the expression profiles of lower-expressed paralogs were less correlated with those of their Brassicaceae orthologs than were those of the higher- and similarly expressed paralogs. This suggests conservation in expression profiles, but not for lower-expressed paralogs, possibly encoded by regulatory DNA sequences [30, 32]. That these differences were statistically not significant, though observed in two independent datasets and three species, might be due to a limited availability of expression data for the non-model organisms.

We used orthologous CNSs as a measure of conservation for the non-coding sequences of paralogous genes. This made it possible to analyze the conservation for each gene of a paralogous pair individually. We found that the lower-expressed genes have fewer CNSs in their promoters, introns, and 3′ UTRs, compared to their higher-expressed paralogs. Thus, we assume that the non-coding sequences of the lower-expressed paralogs have higher mutation rates [54]. The ratio of differential expression was negatively correlated with the number of CNSs for genes with lower expression, suggesting that paralog expression divergence is linked to losses of *cis*-regulatory elements residing in CNSs.

Drivers of paralog retention following the α-WGD event

The retention of duplicated genes following WGD has been explained by several models [14]. Our analysis has revealed that paralogous genes with differential expression are enriched for functions related to responses to different abiotic and biotic stresses. This finding supports the notion that polyploidy is a means to increase adaptability to changing environmental conditions [9], in agreement with the neo- and subfunctionalization models [18]. Indeed, we found that the lower-expressed paralogs of differentially expressed pairs were expressed in a more tissue-specific manner, which has been shown to facilitate neofunctionalization [55].

By contrast, the products of paralogous pairs with highly similar expression levels were enriched for proteins that are subunits of ribosomes. Paralogous genes of *A. thaliana* 80S ribosomal proteins have been shown to be retained by purifying selection and haploinsufficiency [46], indicating that our findings support the notion of gene dosage sensitivity. The balanced gene drive model predicts similar transcript dosage [11, 14]. Accordingly, gene expression profiles must also be similar, and this we found to be the case (Fig. 3).

CNSs possibly function as promoter and enhancer elements

We showed that CNSs are negatively correlated with increasing expression divergence between two paralogs. This effect was most striking for CNSs identified among nine Brassicaceae species (CNS dataset 1), which was also the most closely related group of organisms analyzed. CNSs in promoter regions are enriched for transcription factor binding sites [33, 38, 44, 45] and hence the correlation we observed between CNSs and paralog transcript level ratios could be attributed to DNA elements promoting transcription. Our analysis of TF binding motif enrichment in CNSs identified only two cases that were statistically significant. One of these was the depletion of MYB binding motifs in higher-expressed paralogs. MYB TFs are regulators of processes ranging from primary and secondary metabolism, over cell fate and identity to developmental processes and responses in biotic and abiotic stresses [56], making it difficult to discern the reason for the depletion we found. Similarly-expressed paralogs were depleted for HD-ZIP binding motifs. HD-ZIP TFs participate in organ development and are involved in responses to environmental conditions [57]. This may explain why HD-ZIP binding motifs are depleted at similarly expressed paralogs, which we found to be enriched for functioning as constituents of ribosomes. Notably, the GC content and length of CNSs were similar for all paralogs, suggesting that the significant depletion of TF binding motifs is not an artefact of general sequence differences.

In introns, CNSs are frequently located in regions flanking exons, suggesting a function in splicing regulation

[38], which has been reported to diverge between paralogs in *A. thaliana* [58]. It is more difficult to fathom how the transcript level of a gene is affected by CNSs in the 3′ UTRs. In the datasets available to us, we did not find evidence that mRNA half-lives are influenced by the presence or absence of CNSs in the 3′ UTRs, which has been reported for paralogous CNSs [59]. Alternatively, elements in the 3′ UTRs may act as transcriptional enhancers [60].

Conclusions

Widely accepted models such as the 'balanced gene drive' and 'neo- and subfunctionalization' explain the retention of paralogous genes, but it is not known if these models apply to the coding and non-coding sequences of the same genes. Our data link these models of paralog retention to gene function, coding and non-coding sequences, and gene-expression profiles. Because gene expression profiles are in part established by *cis*-regulatory elements inside CNSs, we propose that similar forces of purifying selection act on coding and non-coding sequences. Taken together, our finding that lower-expressed paralogs have fewer CNSs and are more tissue specific than are higher- or similarly expressed paralogs suggests that CNSs promote or enhance transcription in a broad range of organs or cell types in *A. thaliana*. This is different from findings in metazoans, where CNSs regulate transcription in specific cells [61, 62].

Methods

Sequence data and definitions of gene-associated regions

All data for the *A. thaliana* genome sequence were downloaded from TAIR [63, 64]. Promoter regions were defined as 1500 bp, 1000 bp, and 500 bp upstream of start codons, but omitting ATG -100 bp. 5′ UTRs were defined as the regions 100 bp upstream of the start codon (to rule out any bias for highly expressed genes with well-annotated 5′ UTRs) [65]. Introns encompass all introns of the representative gene model that lie within the protein coding region. 3′ UTRs were defined as the regions 200 bp downstream of stop codons [61]. 3′ downstream regions were defined as the regions 300 to 1000 bp downstream of stop codons.

Gene expression data and measurement of tissue specificity

Robust Multi-array Average (RMA)-normalized microarray data for 79 diverse samples of *Arabidopsis thaliana* tissues and cell types harvested at different developmental stages [47] were downloaded from ArrayExpress [66]. The data were disregarded for probes that detect genes other than those annotated as 'protein-coding' (TAIR10) and probes that hybridize with more than one gene, and also for cases where several probes were annotated as hybridizing to the same gene (according to _at to AGI Conversion Tool [67]). Expression values (obtained in *A. thaliana* accession Col-0) for samples of the same organ were then averaged, namely roots (all samples), green parts of seedlings (7- and 8-days-old), leaves (all samples except data for 'senescing leaves' but including 'cauline leaf'), seeds (all samples, stages 6-10), siliques (all samples, stages 3,4, and 5, with seeds), stem (1st node and 2nd internode), and flowers (all samples, stages 9, 10/11, 12, 15, and one undefined). Data for pedicels, sepals, petals, stamen, carpels (all from flowers at stage 15), hypocotyls, cotyledons, and mature pollen were kept as individual samples (Additional file 1: Table S1). To approximate the overall average gene expression level, the average across the fifteen tissue- or cell type-specific samples was calculated. The tissue specificity for every gene was calculated with the index τ [51]:

$$\tau = \frac{\sum_{i=1}^{N}[1-x_i]}{N-1}$$

where N is the number of tissues and x_i is the expression profile component normalized by the maximal component value. Genes with a τ of close to 0 are broadly expressed across all tissues, while those with tissue-specific expression approach τ = 1.

Paralogous gene pairs

A list of 2563 paralogous gene pairs that evolved after a whole genome duplication (WGD) event 23 Mya was downloaded from a previous publication [6]. Gene pairs that lacked an annotation in TAIR10 for one or both of the genes were excluded. To enable cross-platform analysis of the gene pairs, only those present in the gene expression data (see above) were selected. To strengthen the gene pairs' similarity with respect to the proteins they encode, all pairs with more than 5 % difference in protein coding sequence length (relative to the longer protein; TAIR10) were filtered out. These adjustments resulted in a set of 1312 paralogous pairs that was used for analysis in this study (Additional file 1: Table S4). Paralogous pairs were grouped as similar or differentially expressed when the expression ratio between the two genes of a pair was < 1.25 ($n = 245$) or ≥ 7 ($n = 156$), respectively. For further analysis of the similarly expressed paralogs, one gene per pair was randomly selected, resulting in a list of paralogous genes with "similar expression" (Additional file 1: Table S3). Differentially expressed pairs were divided into sets of "higher expression" or "lower expression".

Substitution rates and orthologs

The BioMart [68] tool in EnsemblPlants [69] was used to retrieve the dN (non-synonymous; change in protein sequence) and dS (synonymous; protein sequence

unchanged) data from genes in *A. lyrata* that are orthologs of the genes in *A. thaliana* described above. Ensembl uses codeml from PAML (Phylogenetic Analysis by Maximum Likelihood) [70] to calculate dN and dS values. The obtained data were cleared for non-representative *A. thaliana* gene models. Where one *A. thaliana* gene had more than one ortholog in *A. lyrata*, only the ortholog with higher '*A. lyrata* % identity' values was retained. If these values were identical, the entry with the lowest dS value was used. Statistical analysis using the Kruskal-Wallis test and Dunn's correction for multiple comparisons was carried out in Prism 6 (GraphPad Software, Inc.). The list of orthologs identified for the *A. thaliana* paralogs was also used to quantify the differentially expressed *A. thaliana* paralogs that had an *A. lyrata* ortholog.

A. lyrata *gene expression data*

Gene expression data for *A. thaliana* and *A. lyrata* orthologs in the form of *Z* scores were provided by the authors [49]. Briefly, for *A. thaliana*, RNA was extracted from whole inflorescences, up to stage 14 flowers, and applied to a tiling array. Triplicate biological samples were generated, with a single technical replicate per sample. For *A. thaliana*, mRNA was extracted from floral tissues (stages 1-14), followed by sequencing (mRNAseq). Two technical replicates each of two biological replicates were sequenced. To make comparisons between the two gene expression datasets, which were identified using different methods, the distribution of expression values was standardized. Data were log-normalized and the resulting normal distribution of gene expression levels was transformed into standard (*Z*) scores, i.e., units of SD from the mean [49].

Statistical analysis (Spearman rank correlation, two-sided) of the correlation between *Z* score expression data for all genes in our analysis and the sets of paralogs with higher expression and lower expression was performed with the function cor.test in R [71, 72]. To compare the Spearman correlation coefficients, ρ was calculated for all orthologs, ρ values were treated as though they were Pearson correlation coefficients (r), and Fisher's z-transformation was used to determine significant differences between two correlation coefficients [73] using VassarStats [74]. Differences in similarly expressed paralogous pairs were further analyzed if *Z* scores were obtainable for both genes. To approximate the correlation between *A. thaliana* and *A. lyrata* gene expression for the similarly expressed genes, an R script (Additional file 2: Script 1) was written that randomly selected expression values from one gene of each paralogous pair and calculated the Spearman correlation coefficient. This was repeated 10,000 times.

GO annotation analysis

Genes from the higher expression and similar expression datasets were analyzed for GO term enrichment using AgriGO [75, 76]. Statistical analysis for GO term enrichment of gene sets was performed using Fisher's exact test with subsequent adjustment for multiple testing by calculating the false discovery rate (FDR) [77]. The minimum number of mapping entries was set to 5, and the list of 19,765 genes derived from the ATH1 microarray was used as a reference.

Conserved non-coding sequence data

Published data for conserved non-coding sequences were downloaded from publicly available websites [78–80]. CNS dataset 1: *A. thaliana* CNSs track was selected and CNSs termed sncCNS were filtered out [38, 78]; CNS dataset 2: mostCons track [39, 79]; CNS dataset 3: BED file of all CNSs [40, 80];.All data files were imported to Google BigQuery [81] and CNSs were retrieved within the 500 bp, 1000 bp, and 1500 bp region upstream of the start codon (ATG -1), in any intron, 200 bp downstream of the stop codon, and within 300 to 1000 bp downstream of the stop codon. Where a neighboring gene extended into the specified upstream or downstream regions (examined using TAIR10 data for intergenic sequences), the queried sequence was shortened accordingly. Only CNSs that were situated entirely within the respective regions were considered for further analyses. If a particular sequence did not contain any CNS, this was counted as zero. Statistical analyses of the correlation (Kendall rank correlation, *tau-b,* which adjusts for tied values) between the number of CNSs (or the sum of bases of all these CNSs) and transcript levels (as $\log_2$-values), or fold-change between two paralogous genes, were performed using the function cor.test in R [71, 72].

Analyses of CNSs

The average length and GC content of CNSs from each CNS dataset and each genetic region (promoter, intronic, or 3′ UTR; as defined above) were tested for significance by resampling 1000 times without replacement; the sample pools were promoter regions of all 19,765 genes in the analysis. Data for transcription factor (TF) binding sites of 274 TFs, belonging to 30 TF families, were obtained from AthaMap [82, 83] and uploaded to Google BigQuery [81]. For further analysis, only those TF binding sites were considered that had a sequence conservation score of at least 50 %. TF binding sites were counted as being inside a CNS (in 1500-bp promoter regions) if the site position started no more than 2 bases upstream of a CNS, or had at least 4 bases overlapping with a CNS. Significance was tested by resampling 1000 times without replacement;

the sample pool was defined as CNSs in promoters of all 19,765 genes in the analysis. Resampling was performed with the Resampling Stats for Excel add-in (statistics.com, LLC).

mRNA half-life data

Data of mRNA half-lives sampled in *A. thaliana* suspension cell cultures were downloaded from the supplemental material accompanying the publication [84]. Paralogous pairs were further analyzed only when half-life data were available for the mRNAs encoded by both genes. Pairs were further selected if one gene lacked a CNS in its 3′ UTR and the other had at least one CNS in its 3′ UTR. Statistical analysis (Wilcoxon signed-rank test) was performed in Prism 6 (GraphPad Software, Inc.).

Additional files

Additional file 1: Table S1. Expression values and TAU (tissue specificity) values of *A. thaliana* genes. Table S2. Paralogous gene pairs in *A. thaliana*; their expression levels and TAU values. Table S3. List of one randomly selected gene from every similarly expressed paralogous pair. Table S4. Normalized expression values of *A. thaliana* and *A. lyrata* orthologs. Table S5. rho values from 10,000 permutations. Table S6. Full list of significantly enriched GO terms. Table S7. Tissue with highest gene expression for differentially expressed paralogous genes. Table S8. CNSs found at *A. thaliana* genes. Table S9. Statistics for the correlation between the number of CNSs at paralogous genes and expression levels. Table S10. mRNA half-life data for genes included in this study. Table S11. mRNA half-life data for paralogous pairs that have 3′ UTR CNSs in one gene only. (XLSX 6431 kb)

Additional file 2: Figure S1. Correlation between gene expression levels of orthologs in *A. thaliana* and *A. lyrata*. Correlation between gene expression in the lower expression (a) and higher expression (b) data sets measured in *A. thaliana* and *A. lyrata*. Figure S2. Correlation between gene expression levels of orthologs in four Brassicaceae species. Correlation between gene expression in the higher expression (a-c) and lower expression (d-f) data sets measured in *A. thaliana* and *A. lyrata* (a, d), *A. thaliana* and *C. rubella* (b, e), and *A. thaliana* and *C. grandiflora* (c, f). Figure S3. Correlation between differential gene expression and the number of conserved nucleotides in the promoters, introns, and 3′ UTRs of paralogous pairs. The number of conserved nucleotides in CNSs within the 1500-bp upstream regions (a, d, g), introns (b, e, h), or 3′ UTRs (c, f, i) of all genes present on the ATH1 microarray was determined using three different CNS datasets. For every gene, the average transcription level (log_2-value) is given on the y-axis. For lower-expressed paralogs, CNSs and transcription levels are negatively correlated (a, b, c, f, and g). Script 1: R script to calculate rho-values. (PDF 490 kb)

Acknowledgements

We thank Jesse D Hollister and Brandon S Gaut for sharing processed *A. thaliana* and *A. lyrata* gene expression data, Adrian E Platts and Stephen I Wright for sharing processed *A. thaliana*, *A. lyrata*, *C. rubella*, and *C. grandiflora* gene expression data, Matteo Scortichini for writing R scripts, Christian Storm Pedersen for valuable comments, and Kathleen Farquharson for language editing and valuable comments.

Funding

This work was supported by the Danish National Research Foundation through the PUMPkin Centre of Excellence.

Availability of data and material

The data sets supporting the conclusions of this article are included within the article and its additional files.

Authors' contributions

RDH and MP conceived the study and designed experiments. RDH carried out experiments and data analyses. RDH and MP wrote the manuscript and approved its final version.

Competing interests

The authors declare that they have no competing interests.

Received: 25 November 2015 Accepted: 27 May 2016
Published online: 13 June 2016

References

1. Blanc G, Hokamp K, Wolfe KH. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. 2003; 13(2):137–44.
2. Friedman R, Hughes AL. Pattern and timing of gene duplication in animal genomes. Genome Res. 2001;11(11):1842–7.
3. McLysaght A, Hokamp K, Wolfe KH. Extensive genomic duplication during early chordate evolution. Nat Genet. 2002;31(2):200–4.
4. Barker MS, Vogel H, Schranz ME. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. Genome Biol Evol. 2009;1:391–9.
5. Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y. The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 2002;99(21):13627–32.
6. Knowles DG, McLysaght A. High rate of recent intron gain and loss in simultaneously duplicated *Arabidopsis* genes. Mol Biol Evol. 2006;23(8):1548–57.
7. Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, et al. The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci U S A. 2015;112(27):8362–6.
8. Lockton S, Gaut BS. Plant conserved non-coding sequences and paralogue evolution. Trends Genet. 2005;21(1):60–5.
9. Ohno S. Evolution by gene duplication. Berlin Heidelberg: Springer; 1970.
10. Zhang JZ. Evolution by gene duplication: an update. Trends Ecol Evol. 2003;18(6):292–8.
11. Freeling M, Thomas BC. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 2006; 16(7):805–14.
12. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151(4):1531–45.
13. Lewis EB. Pseudoallelism and gene evolution. Cold Spring Harb Symp Quant Biol. 1951;16:159–74.
14. Freeling M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 2009;60:433–53.
15. Casanova-Saéz R, Candela H, Micol JL. Combined haploinsufficiency and purifying selection drive retention of *RPL36a* paralogs in Arabidopsis. Sci Rep. 2014;4:4122.
16. Birchler JA, Riddle NC, Auger DL, Veitia RA. Dosage balance in gene regulation: biological implications. Trends Genet. 2005;21(4):219–26.
17. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000;154(1):459–73.
18. Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell. 2004;16(7):1679–91.
19. Ha M, Kim ED, Chen ZJ. Duplicate genes increase expression diversity in closely related species and allopolyploids. Proc Natl Acad Sci U S A. 2009; 106(7):2295–300.
20. Harikrishnan SL, Pucholt P, Berlin S. Sequence and gene expression evolution of paralogous genes in willows. Sci Rep. 2015;5:18662.
21. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006;444(7116):171–8.
22. McGrath CL, Gout JF, Johri P, Doak TG, Lynch M. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. Genome Res. 2014;24(10):1665–75.
23. Papp B, Pal C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. Nature. 2003;424(6945):194–7.
24. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, et al. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 2005;102(15):5454–9.

25. Seoighe C, Gehring C. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends Genet. 2004; 20(10):461–4.
26. Brunet FG, Roest Crollius H, Paris M, Aury JM, Gibert P, Jaillon O, et al. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol Biol Evol. 2006;23(9):1808–16.
27. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 2006;7(5):R43.
28. Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, et al. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. Genome Res. 2012;22(1):95–105.
29. Geiser C, Mandakova T, Arrigo N, Lysak MA, Parisod C. Repeated whole-genome duplication, karyotype reshuffling and biased retention of stress-responding genes in Buckler Mustards. Plant Cell. 2016;28(1):17-27.
30. Richardson DN, Wiehe T. Properties of sequence conservation in upstream regulatory and protein coding sequences among paralogs in *Arabidopsis thaliana*. In: Ciccarelli FD, Miklós I, editors. Comparative Genomics. Berlin Heidelberg: Springer; 2009. p. 217–28.
31. Spangler JB, Subramaniam S, Freeling M, Feltus FA. Evidence of function for conserved noncoding sequences in *Arabidopsis thaliana*. New Phytol. 2012; 193(1):241–52.
32. Haberer G, Hindemitt T, Meyers BC, Mayer KF. Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. Plant Physiol. 2004;136(2):3009–22.
33. Freeling M, Subramaniam S. Conserved noncoding sequences (CNSs) in higher plants. Curr Opin Plant Biol. 2009;12(2):126–32.
34. Nelson AC, Wardle FC. Conserved non-coding elements and *cis* regulation: actions speak louder than words. Development. 2013;140(7):1385–95.
35. Hardison RC. Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. 2000;16(9):369–72.
36. Baxter L, Jironkin A, Hickman R, Moore J, Barrington C, Krusche P, et al. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. Plant Cell. 2012;24(10):3949–65.
37. Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C, Rubin EM, et al. Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res. 2000;10(9):1304–6.
38. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. 2013;45(8):891–8.
39. Hupalo D, Kern AD. Conservation and functional element discovery in 20 angiosperm plant genomes. Mol Biol Evol. 2013;30(7):1729–44.
40. Van de Velde J, Heyndrickx KS, Vandepoele K. Inference of transcriptional networks in *Arabidopsis* through conserved noncoding sequence analysis. Plant Cell. 2014;26(7):2729–45.
41. Hettiarachchi N, Kryukov K, Sumiyama K, Saitou N. Lineage-specific conserved noncoding sequences of plant genomes: their possible role in nucleosome positioning. Genome Biol Evol. 2014;6(9):2527–42.
42. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 2005;3(1):e7.
43. Burgess D, Freeling M. The most deeply conserved noncoding sequences in plants serve similar functions to those in vertebrates despite large differences in evolutionary rates. Plant Cell. 2014;26(3):946–61.
44. Freeling M, Rapaka L, Lyons E, Pedersen B, Thomas BC. G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in *Arabidopsis*. Plant Cell. 2007;19(5):1441–57.
45. Levy S, Hannenhalli S, Workman C. Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics. 2001;17(10):871–7.
46. Burgess DG, Xu J, Freeling M. Advances in understanding *cis* regulation of the plant gene with an emphasis on comparative genomics. Curr Opin Plant Biol. 2015;27:141–7.
47. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, et al. A gene expression map of *Arabidopsis thaliana* development. Nat Genet. 2005;37(5):501–6.
48. Hu T, Pattyn P, Bakker E, Cao J, Cheng J-F, Clark R, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 2011;43(5):476–81.
49. Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc Natl Acad Sci U S A. 2011;108(6):2322–7.
50. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet. 2013;45(7):831–5.
51. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005; 21(5):650–9.
52. Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. Analysis of sequence conservation at nucleotide resolution. PLoS Comput Biol. 2007; 3(12):e254.
53. Harmston N, Baresic A, Lenhard B. The mystery of extreme non-coding conservation. Philos Trans R Soc Lond B Biol Sci. 2013;368(1632):20130021.
54. Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. PLoS Genet. 2009;5(7):e1000581.
55. Thompson A, Vo D, Comfort C, Zakon HH. Expression evolution facilitated the convergent neofunctionalization of a sodium channel gene. Mol Biol Evol. 2014;31(8):1941–55.
56. Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L. MYB transcription factors in *Arabidopsis*. Trends Plant Sci. 2010;15(10):573–81.
57. Ariel FD, Manavella PA, Dezar CA, Chan RL. The true story of the HD-Zip family. Trends Plant Sci. 2007;12(9):419–26.
58. Tack DC, Pitchers WR, Adams KL. Transcriptome analysis indicates considerable divergence in alternative splicing between duplicated genes in *Arabidopsis thaliana*. Genetics. 2014;198(4):1473–81.
59. Spangler JB, Feltus FA. Conserved non-coding sequences are associated with rates of mRNA decay in *Arabidopsis*. Front Plant Sci. 2013;4:129.
60. Jash A, Yun K, Sahoo A, So JS, Im SH. Looping mediated interaction between the promoter and 3′ UTR regulates type II collagen expression in chondrocytes. PLoS ONE. 2012;7(7):e40828.
61. Zheng Y, Josefowicz S, Chaudhry A, Peng XP, Forbush K, Rudensky AY. Role of conserved non-coding DNA elements in the *Foxp3* gene in regulatory T-cell fate. Nature. 2010;463(7282):808–12.
62. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006;444(7118):499–502.
63. The Arabidopsis Information Resource. ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release. Accessed 23 February 2016.
64. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, et al. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Res. 2003;31(1):224–8.
65. Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA. Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. Plant Mol Biol. 2006;60(1):69–85.
66. ArrayExpress. http://www.ebi.ac.uk/arrayexpress/files/E-TABM-17. Accessed 23 Feb 2016.
67. _at to AGI Conversion Tool. http://bbc.botany.utoronto.ca/ntools/cgi-bin/ntools_agi_converter.cgi. Accessed 23 Feb 2016.
68. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015;43(W1):W589-98.
69. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):9.
70. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.
71. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015.
72. The R Project for Statistical Computing. https://www.r-project.org. Accessed 23 Feb 2016.
73. Myers L, Sirois MJ. Spearman correlation coefficients, Differences between. In: Kotz S, editor. Encyclopedia of Statistical Sciences. New York: John Wiley and Sons; 2006.
74. VassarStats: Website for Statistical Computation. vassarstats.net/rdiff.html. Accessed 23 Feb. 2016.
75. GO Analysis Toolkit and Database for Agricultural Community. http://bioinfo.cau.edu.cn/agriGO/. Accessed 23 Feb 2016.
76. Yi X, Du Z, Su Z. PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res. 2013;41(Web Server issue):W98–103.

77. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J Roy Stat Soc B Met. 1995;57(1):289–300.
78. UCSC Genome Browse. http://mustang.biol.mcgill.ca:8885/cgi-bin/hgTables. Accessed 23 Feb 2016.
79. A. thaliana Genome Browser. http://genome.genetics.rutgers.edu/cgi-bin/hgGateway?hgsid=2310&clade=plant&org=0&db=0. Accessed 23 Feb 2016.
80. Arabidopsis conserved non-coding sequences supporting data. http://bioinformatics.psb.ugent.be/cig_data/Ath_CNS/AllFootPrintsFDR0.10_scores.bed. Accessed 23 Feb 2016.
81. Google BigQuery. https://developers.google.com/bigquery. Accessed 23 Feb 2016.
82. Hehl R, Norval L, Romanov A, Bülow L. Boosting AthaMap database content with data from protein binding microarrays. Plant Cell Physiol. 2016;57(1):e4.
83. AthaMap. http://www.athamap.de. Accessed 23 Feb 2016.
84. Narsai R, Howell KA, Millar AH, O'Toole N, Small I, Whelan J. Genome-wide analysis of mRNA decay rates and their determinants in *Arabidopsis thaliana*. Plant Cell. 2007;19(11):3418–36.